Plan de gestion de données ANR
Recommandations de l'Université de Lille

Le présent document a été conçu pour aider les coordinateurs de projets financés à remplir le plan de gestion de données : éléments pouvant figurer dans les différentes parties du modèle ANR, contacts et liens utiles.

Il compile les recommandations de l'ANR et les recommandations des services compétents de l'Université de Lille : Service Commun de la Documentation (SCD), Direction Données Personnelles et Archives (DDPA), Direction du Numérique.

Ce modèle de PGD s'appuie sur la trame de l'ANR[1] mais les recommandations sont valables, quelle que soit l'agence de financement.

Pour bénéficier de conseils personnalisés, d'appui pour sélectionner des outils et/ou d'une relecture commentée du PGD, vous pouvez contacter l'atelier de la donnée Lille open research data : bit.ly/lord-support

Pour plus d'informations, consulter la page : https://bu.univ-lille.fr/chercheurs-doctorants/science-ouverte/projets-finances

[1] Modèle de PGD Structuré (2023) : https://dmp.opidor.fr/public_templates

Informations générales

Renseignements administratifs :

Titre du projet :

Acronyme du projet :

Date de début et de fin du projet :

Nom et prénom du coordinateur :

Affiliation :

Contact concernant le PGD :

Version du PGD :

Date du PGD :

Courte description

Une très courte description du projet permettra au lecteur de mieux comprendre le contexte du PGD.

Versionnage :

exemple :

Phase du PGD	Numéro de version du PGD	Date	Description	Auteurs
initial	v.1	21/04/2024	Premier brouillon	Personne1
initial	v.1.1	15/05/2024	Relecture et mises à jour mineures	Personne2
finale	v.2	04/09/2024	Version remise au financeur	Personne1

1. Description des données et collecte ou réutilisation de données existantes

1.1 Description générale du produit de recherche

- Donner des détails sur le **type de données**
    - par exemple : caractéristiques d'échantillons biologiques, code source logiciel, corpus d'archives, textes, entretiens, imagerie, bases de données d'enquêtes, résultats d'expériences et de mesures, données de carottage, séquences génétiques, données sismographiques, enregistrement audio ou vidéo, photographies, relevés météo, mesures sismiques, statistiques de population, données de simulation numérique (modèles climatiques), tests en psychologie, séquences de peptides…
- Détailler le **format** attendu des données : la manière selon laquelle les données sont codées pour le stockage, généralement reflétée par l'extension du nom de fichier (par exemple PDF, XLS, doc, TXT, ou RFD).
    - Justifier l'utilisation de certains formats. Par exemple, les choix d'un format peuvent être guidés par l'expertise du personnel de l'organisme, ou par une préférence pour les formats ouverts, par les standards de format acceptés par les entrepôts de données, par l'usage largement répandu dans une communauté de recherche ou par le logiciel ou l'équipement qui seront utilisés.
    - Privilégier les formats standards et ouverts (exemple : CSV plutôt que XLS), car ils facilitent le partage et la réutilisation à long terme des données (plusieurs catalogues fournissent des listes de ces « formats préférés »).
    - L'outil FACILE du CINES permet de vérifier que les formats utilisés sont pérennes et archivables.
    - Le Référentiel général d'interopérabilité (RGI) donne accès à des recommandations référençant des normes et standards qui favorisent l'interopérabilité au sein des systèmes d'information de l'administration.

Sur les enjeux liés aux formats, voir : https://doranum.fr/stockage-archivage/quiz-format-ouvert-ou-ferme_10_13143_mcwq-qs64/

- Détailler les **volumes** (qui peuvent être exprimés en espace de stockage requis [octets], et/ou en quantités d'objets, de fichiers, de lignes, et colonnes).
    - Dans la version initiale du PGD, la volumétrie demandée n'est qu'une estimation qui pourra être réévaluée lors des versions ultérieures.

1.2 Est-ce que des données existantes seront réutilisées ?

- Si non : énoncer les éventuelles contraintes à la réutilisation des données préexistantes. Indiquer les raisons pour lesquelles l'utilisation de sources de données existantes a été envisagée mais écartée.

- Si oui : indiquer la source des données réutilisées ainsi que les droits qui y sont associés (via des licences par exemple).

Nota Bene RGPD : La réutilisation des données personnelles issues d'une recherche est possible, à condition que le responsable scientifique respecte ces dispositions : contacter le DPO qui doit faire la déclaration du nouveau projet ; s'assurer que les participants ont été informés de la possible réutilisation des données lors de la collecte initiale, et les informer du nouveau projet (car elles doivent pouvoir s'opposer à la réutilisation de leurs données, si elles le souhaitent) ; la réutilisation a pour objectif de répondre à une question précise et ponctuelle ; la durée de recherche est limitée et connue.

1.3 Comment seront produites / collectées les nouvelles données ?

Pour les données produites *ex-nihilo* :

- Expliquer quelles méthodologies ou quels logiciels seront utilisés pour la production/collecte de données (nombre d'entretiens envisagés, support d'enregistrement des réponses, etc.).
- Expliquer comment la provenance des données sera documentée.

2. Documentation et qualité des données

2.1 Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

- **Métadonnées** : données décrivant d'autres données qui permettent d'identifier et de comprendre la ressource décrite.
    - Indiquer quelles métadonnées seront fournies pour aider à la recherche et à l'identification des données. Préciser leur type : métadonnées externes que l'on ajoute soi-même afin de décrire les données, et/ou métadonnées embarquées qui sont générées automatiquement lors de la création des données.
    - Indiquer quels **schémas** ou **standards de métadonnées** seront utilisés

      Schéma : modèle qui fournit un ensemble d'éléments caractéristiques qui permettent de décrire les données. Le catalogue **FAIRsharing** permet de trouver un schéma de métadonnées adapté.

      Standard : modèle de description reconnu dans une communauté ou une discipline afin d'homogénéiser les pratiques des chercheurs. Utiliser les standards de métadonnées des communautés scientifiques lorsque ceux-ci existent (par exemple DDI, TEI, EML, MARC, CMDI).

- Penser à la **documentation** qui serait nécessaire pour permettre la compréhension, garder la mémoire du traitement des données et assurer leur réutilisation. Il peut s'agir notamment de l'information sur la méthodologie utilisée pour collecter les données, sur les procédures et méthodes d'analyse utilisées, sur la définition des variables, des unités de mesure, etc.

    - Mentionner la façon dont ces informations seront obtenues et enregistrées par exemple dans une base de données avec des liens vers chacun des fichiers, dans un fichier texte de type « lisez-moi »/« readme », dans les en-têtes de fichiers, dans un livre de référence (« code book ») ou dans les cahiers de laboratoire.

    - Les fichiers *readme* sont des documents textes, compilant les métadonnées essentielles à la compréhension, ajoutés aux jeux de données lors du dépôt dans un entrepôt. C'est un guide ayant pour but d'aider les chercheurs à comprendre les ensembles de données et de quelle manière les réutiliser.

    - Indiquer si des vocabulaires contrôlés ou des ontologies spécifiques sont utilisés pour les données et si oui, lesquels.

- **Organisation des données :** indiquer comment les données seront organisées au cours du projet, en mentionnant par exemple les conventions de nommage, le contrôle de version et les structures des dossiers. Des données bien classées et gérées de façon cohérente seront plus faciles à retrouver, à comprendre et à réutiliser.

    - **Structure des dossiers :** Séparer les données brutes des données traitées. **Exemple** : L'arborescence peut être thématique (gouvernance, expérimentation, publication…). Éviter de créer trop de sous-dossiers afin de limiter le nombre de clics pour accéder à l'information.

    - **Convention de nommage :** Permet d'expliciter et d'homogénéiser les intitulés des fichiers. Respecter la même organisation facilite le traitement automatique des fichiers dans un système informatique.

    - **Versioning :** Utiliser des outils collaboratifs disposant d'un système de conservation de l'historique. Cela favorise aussi la récupération de versions antérieures en cas d'incidents. L'outil Nextcloud de l'université de Lille permet cette gestion. Toujours modifier des fichiers en ayant activé l'option «*suivi des modifications*».

Voir le guide « Organiser et nommer ses documents numériques » (ENT : https://ent.univ-lille.fr/environnement-de-travail/archives)

2.2 Quelles seront les mesures de contrôle utilisées pour assurer la qualité scientifique des données ?

- Expliquer comment la qualité et la conformité de la collecte des données seront contrôlées et documentées. Il s'agit là de préciser les processus comme la calibration, la répétition des

échantillons ou des mesures, la capture standardisée des données, la validation de saisie des données, la revue par les pairs, ou la représentation basée sur des vocabulaires contrôlés.

3. Exigences légales et éthiques, codes de conduite

3.1 Quelles seront les mesures appliquées pour assurer la protection des données à caractère personnel ?

- Si des données à caractère personnel sont utilisées dans la recherche, il est obligatoire de contacter le **DPO** (délégué à la protection des données) de l'université de Lille (dpo@univ-lille.fr). Le **référent RGPD** (Règlement Général sur la Protection des Données) du laboratoire peut être contacté également.

- Si des données à caractère personnel sont utilisées dans la recherche, veiller à ce que les lois sur la protection des données (par exemple, RGPD) soient appliquées, notamment :

    - Obtenir un consentement éclairé pour la collecte, la préservation et/ou le partage de données personnelles. Si un recueil de consentement a été établi, toujours en garder une trace.

    - Envisager l'anonymisation des données personnelles pour la préservation et/ou le partage (des données correctement anonymisées ne sont plus considérées comme des données personnelles). L'outil Amnesia permet d'anonymiser des données.

    - Envisager la pseudonymisation des données personnelles (la principale différence avec l'anonymisation est que la pseudonymisation est réversible).

    - Envisager le chiffrement des données, qui est considéré comme un cas particulier de pseudonymisation (la clé de cryptage doit alors être stockée séparément des données, par exemple par un tiers de confiance). L'outil VeraCrypt permet de crypter des données.

    - Expliquer si une procédure d'accès spécifique a été mise en place pour les utilisateurs autorisés à accéder aux données personnelles.

3.2 Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

- Expliquer qui sera le propriétaire des données, qui aura le droit d'en contrôler l'accès.

- Quelles conditions d'accès s'appliqueront aux données ? Les données seront-elles librement accessibles, ou des restrictions seront-elles appliquées ? Si oui, lesquelles ?

- Envisager l'utilisation de licences concernant l'accès et la réutilisation des données. Favoriser l'utilisation de licences ouvertes Creative Commons ou Etalab pour les données de recherche.

  Plus d'informations https://bu.univ-lille.fr/chercheurs-doctorants/science-ouverte/donnees-de-recherche/introduction-a-la-propriete-intellectuelle-pour-la-gestion-des-donnees.

- S'assurer de couvrir, dans l'accord de consortium, ces questions de droits de contrôle d'accès aux données pour les projets multipartenaires et en cas de propriété partagée des données. Reprendre les points qui y sont abordés en précisant à qui appartiennent les données produites et collectées, ainsi que les droits régissant les différentes bases de données impliquées le cas échéant.

Nota Bene : pour des données créées ex-nihilo et dans la grande majorité des cas, les données produites sont publiques (pas de droit d'auteur par défaut) et la propriété en revient à l'établissement de tutelle. Cela implique un droit d'accès sur demande, une diffusion gratuite et une libre réutilisation.

- Indiquer si les droits de propriété intellectuelle (par exemple la directive bases de données, droits sui generis) sont affectés. Dans l'affirmative, expliquer lesquels et comment cela sera traité.

- Indiquer s'il y a des restrictions sur la réutilisation des données fournies par des tiers.

3.3 Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

- Dans le cas de données impliquant la personne humaine, contacter le Comité d'éthique pour la recherche. Dans le cas de données issues d'expériences biomédicales sur l'être humain, contacter le Comité de protection des personnes.

  - Si le CER ou le CPP ont été consultés : préciser lequel, à quelle date et les conclusions de ce dernier.

  Plus d'informations sur le CER : https://www.univ-lille.fr/recherche/la-recherche-au-service-de-la-societe/comprendre-notre-demarche-ethique. + sur l'intranet Recherche

  Plus d'informations sur http://www.comite-de-protection-des-personnes-nord-ouest-iv-lille.sitew.fr/.

- Déterminer si les questions d'éthique auront une incidence sur la façon dont les données seront stockées et transférées, qui pourra les voir ou les utiliser et quelles durées de conservation leur seront appliquées. Démontrer que ces aspects sont bien pris en compte et planifiés. Si oui, indiquer si un comité éthique a été consulté : lequel, à quelle date, conclusions et préconisations.

- Adopter les codes de conduite nationaux et internationaux et le code d'éthique institutionnel et vérifier si une revue des pratiques (par exemple par un comité d'éthique) est requise pour ce qui concerne la collecte de données dans le cadre du projet de recherche.

- Préciser si des autorisations et protocoles spécifiques à certaines disciplines et certains objets de recherche sont requis.

4. Traitement et analyse des données

- Comment et avec quels moyens sont traitées les données ?

5. Stockage et sauvegarde des données pendant le processus de recherche

5.1 Comment les données seront-elles stockées et sauvegardées tout au long du projet ?

- Décrire l'endroit où les données seront stockées et sauvegardées au cours du processus de recherche et la fréquence à laquelle la sauvegarde sera effectuée.
    - **Bonnes pratiques :**
        - Privilégier l'utilisation de systèmes de stockage robustes, avec sauvegarde automatique, tels que ceux fournis par l'université (cf *infra*).
        - Le stockage des données sur des ordinateurs portables, des disques durs externes, ou des périphériques de stockage tels que des clés USB n'est pas recommandé.
        - Règle du 3-2-1 : 3 copies des données sur au moins 2 supports différents, dont au moins 1 copie à distance avec synchronisation.
        - Adapter la fréquence des sauvegardes en fonction de l'état d'avancement du projet : accentuer la fréquence lors de phases fortement changeantes telle que celle de la collecte.
- Expliquer qui aura accès aux données au cours du processus de recherche et comment l'accès aux données sera contrôlé, en particulier dans le cadre de recherches impliquant plusieurs chercheurs / équipes de recherche.

- **Exemple** : protéger l'accès aux données via des rôles, des droits et des mots de passe (possible avec les outils de travail collaboratif).

- Tenir compte de la protection des données, en particulier si vos données sont sensibles (par exemple données à caractère personnel, sensibles, secrets commerciaux). Expliquer quelle politique institutionnelle de protection des données est mis en œuvre.
- Expliquer comment les données seront récupérées en cas d'incident. Mettre en place des procédures de récupération fiables et qui soient testées par plusieurs membres de l'équipe.

Plusieurs solutions logicielles sont apportées par l'université de Lille. L'université garantit la sécurité des données stockées ainsi que le respect des règlementations, en particulier en ce qui concerne la protection des données personnelles :

- **Nextcloud**, dans une limite initiale de 50 Go. Il est possible de demander un agrandissement de l'espace disponible auprès de la DGDNum (Direction Générale Déléguée au Numérique) ;
- **Serveurs virtuels de grande capacité** jusqu'à 10 To ;
- Le **Mésocentre de Calcul Scientifique Intensif** permet d'effectuer de la simulation et de la modélisation de données. Les capacités de stockage et de traitement sont paramétrables auprès de la DGDNum ;
- **Filesender**, pour le partage sécurisé et chiffré de fichiers lourds jusqu'à 100 Go gratuitement.

Les GAFAM en général et Google Drive en particulier sont déconseillés par la CNIL dans l'Enseignement supérieur, car ils n'assurent pas la sécurité des données et la conservation des droits sur les données.

Plus d'information : https://bu.univ-lille.fr/chercheurs-doctorants/science-ouverte/donnees-de-recherche/sauvegarde-et-stockage-des-donnees

6. Partage des données et conservation à long terme

Le partage, la diffusion des données et la conservation à long terme concernent la durée de vie des données après le projet. Dans la version initiale du PGD, vous devrez anticiper, émettre des suppositions d'entrepôts de données et de délais. Il sera ensuite possible d'affiner les réponses au fil des différentes versions du PGD.

6.1 Comment les données seront-elles partagées ?

- Expliquer comment les données pourront être retrouvées et partagées (par exemple, par le dépôt dans un entrepôt de données de confiance, l'indexation dans un catalogue, par l'utilisation d'un service de données sécurisé, par le traitement direct des demandes de données, ou l'utilisation de tout autre mécanisme).

- Indiquer où les données seront déposées. Si aucun entrepôt reconnu n'est proposé, démontrer dans le PGD que les données pourront être prises en charge efficacement au-delà de la durée de financement du projet.
    - Privilégier le dépôt dans un entrepôt de données de confiance (assurant sécurité et préservation, critères consultables sur Ouvrir la Science).
        - CoreTrustSeal propose une liste d'entrepôts certifiés.
        - Re3data permet de trouver un entrepôt disciplinaire ou pluridisciplinaire s'il n'y a pas de recommandation claire dans le laboratoire ou par les partenaires.
        - Le Comité pour la science ouverte propose également une première liste d'entrepôts thématiques, associée à une note méthodologique.
    - En l'absence d'entrepôt disciplinaire reconnu, l'université conseille de déposer les données sur Lillodata.
    - Il est recommandé de justifier le choix de l'entrepôt (présence d'un DOI, utilisation de standards de métadonnées, procédures de dépôt, gratuité du dépôt et de l'accès…).
- Expliquer à quel moment les données seront rendues disponibles. Indiquer les délais de publication prévus. Expliquer si une utilisation exclusive des données est revendiquée et, dans l'affirmative, pour quelle raison et pour combien de temps. Indiquer si le partage des données sera différé ou limité, par exemple pour des raisons de publication, pour protéger la propriété intellectuelle ou le dépôt de brevets.
- Indiquer qui pourra utiliser les données et si une licence d'utilisation sera associée au jeu de données. S'il s'avère nécessaire de restreindre l'accès pour certaines communautés ou d'imposer un accord pour le partage de données, expliquer comment et pourquoi. Expliquer les mesures qui seront prises pour dépasser ou minimiser ces restrictions.

Exemples d'exceptions permettant le non partage des données : les données uniquement communicables à l'intéressé (vie privée, secret médical), secret des affaires, les documents réalisés en exécution d'un contrat de prestation de services exécuté pour le compte d'une ou plusieurs personnes déterminées, les données couvertes par des droits de propriété intellectuelle, etc.

Voir par exemple le Guide d'application de la loi pour une république numérique de 2016 : https://www.ouvrirlascience.fr/guide-application-loi-republique-numerique-article-30-ecrits-scientifiques-version-courte/

- Décrire les utilisations (et/ou les utilisateurs) prévisibles des données dans un cadre de recherche (potentiel de réutilisation du jeu de données).
    - Montrer que la question de la réutilisation des données a été réfléchie : déterminer dans quel contexte les données pourraient être utiles à d'autres chercheurs dans le futur, que ce soit dans la même discipline ou une autre.

- Les identifiants pérennes devraient être appliqués de manière à ce que les données puissent être localisées et référencées de façon fiable et efficace. Les identifiants pérennes aident aussi à comptabiliser les citations et les réutilisations.
    - Indiquer s'il sera envisagé d'attribuer aux données un identifiant pérenne. Typiquement, un entrepôt de confiance attribuera automatiquement des identifiants pérennes. C'est le cas par exemple de Recherche Data Gouv.
    - Lorsque les données sont mises à disposition sur une base de données qui ne dispose pas d'une attribution de DOI automatisée, l'université de Lille peut mettre en place l'attribution de DOI, via une convention entre le SCD et DataCite (contact : science-ouverte@univ-lille.fr).
- Indiquez si les utilisateurs potentiels ont besoin d'outils spécifiques pour l'accès et la (ré)utilisation des données. Tenez compte de la durée de vie des logiciels nécessaires pour accéder aux données.

6.2 Comment les données seront-elles conservées à long terme ?

- Définir le plan de préservation des données et fournir l'information sur la durée d'archivage des données.
    - **Attention**, si une conservation des données au-delà de 10 ans est envisagée, il faudra prévoir un archivage pérenne. Cela nécessite la mise en place d'un processus précis et complexe nécessitant un savoir-faire spécifique.
    - Pour toute question, contacter le service des archives de l'Université : archives@univ-lille.fr.
- Indiquer si des données ne peuvent pas être divulguées ou peuvent être détruites pour des raisons contractuelles, légales, ou réglementaires.
- Indiquer quelles données conserver, en mettant en évidence les critères de sélection des données pour la préservation à long/moyen terme.
    - Les données sont à conserver à minima la durée du projet. Certains documents seront à conserver plus longtemps. Par exemple, les cahiers de laboratoire sont à conserver 25 ans.
    - Il convient de s'interroger sur l'intérêt d'une conservation plus longue des données, en lien avec le DPO et le service archives, selon les possibilités de leur réutilisation, le besoin éventuel d'anonymiser les données personnelles et la nécessité de garantir les preuves de l'intégrité scientifique du projet de recherche.

Sur l'archivage des données, voir :

https://www.cnil.fr/fr/passer-laction/les-durees-de-conservation-des-donnees